DOCUMENT RESUME

ED 254 534 TM 850 034

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE Bias in Prediction: A Closer Look at Theory, Findings, and Conclusions.
PUB DATE 84
NOTE 44p.
PUB TYPE Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Research; Groups; *Predictive Validity; Regression (Statistics); *Research Methodology; Scores; *Test Bias; Test Theory

ABSTRACT

The findings and conclusions from research on predictive validity among markedly different groups are discussed. Empirical findings using the regression line neither support nor contradict the existence of bias or the hypothesis of differential validity. Conclusions drawn from the research are questionable. Elimination of bias in a test, contrary to published reports, is shown to be beneficial to the low scoring group. As a result of these considerations, it has been determined that the main issue is bias in the test, and the related issues of differential validity and differential prediction cannot be addressed until bias has been convincingly detected and eliminated from the test. Ten tables and figures illustrating the findings are appended. A 12-item bibliography is also included. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED254534

BIAS IN PREDICTION:

A CLOSER LOOK AT THEORY, FINDINGS, AND CONCLUSIONS

Eulogio Gonzalez-Tamayo

Azul, Inc. Business Training Center

New York, 1984

TM 850 034

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E. Gonzalez-Tamayo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SUMMARY

The general findings and conclusions from research on predictive validity among markedly different groups are discussed in relation to a theoretical framework of bias. The empirical findings using the regression line neither support nor contradict the existence of bias or the hypothesis of differential validity. The conclusions from the empirical findings are questionable. Elimination of bias in the test, contrary to published reports, is beneficial to the low scoring group.

INTRODUCTION

The issues of bias and fairness in aptitude testing have attracted considerable public and professional attention during the last twenty years. The purpose of any test is to differentiate among individuals; a test that does not differentiate is useless. The question is whether tests can also differentiate among groups markedly different in background and, assuming that they do, the problem is to determine whether the found differences are real or they also represent, at least to some extent, an artifact of the psychometric techniques used.

Assuming that different groups may also be different in ability, researchers have tried to separate psychometrically detectable bias in the test from possible differences on the trait but the efforts have not been completely satisfactory (Ironson, 1982). Differences in the factorial nature of the test, differences in the item parameters and in the test predictive power (validity coefficients and regression equations) among ethnic groups have been studied as possible indicators of bias. The concentration in this paper is on the differences in prediction among different groups as indicators of possible bias in prediction. The central issue in this respect is differential prediction rather than single-group or differential validity (Linn, 1978).

An enormous number of studies have been conducted about the validity coefficient and the regression lines for different groups and the findings have been quite consistent across groups, across tests, and across uses of tests. There is broad agreement among researchers about the findings of these studies; in the final analysis, findings represent facts and no fair minded person can argue with facts.

1. Empirical data have consistently been unable to prove the existence of single-group or differential validity. Correlation coefficients between ability tests and future performance in school or on the job are somewhat lower for minority than for majority groups but the coefficients are comparable if not equal.

2. The analysis done with the regresion lines for minority and majority groups have consistently been unable to prove any systematic over- or underprediction in

either group.

3. When the majority or the common regression line is used, it results in overprediction of average minority criterion scores. The common regression line underpredicts the average majority criterion score.

4. When the two regression lines do differ, almost without exception, the difference is in the intercept being it somewhat higher for the majority than for the minority group.

Some conclusions have been proposed from these general findings which are not supported by every researcher. A conclusion is an interpretation of facts and, although the interpretation may seem reasonable, it goes beyond the facts and fair minded people may draw different conclusions from the same facts.

1. Aptitude tests are not biased agains minority groups (Wigdor & Garner, 1982), at least not agains any native-born English speaking minority group (Jensen, 1980, 1984; Hunter et Al., 1984).

2. Aptitude tests are equaly valid for minority and majority groups (Wigdor & Garner, 1982; Linn, 1978; Boehm, 1977).

3. If there is bias in the test, it will favor the minority groups (Jensen, 1980, 1984; Hunter et Al., 1984) and the elimination of bias will result in the selection of fewer minority persons.

These findings and conclusions seem to settle the controversy about predictive bias in aptitude testing once and for all. Not quite so. The difference between findings and conclusions is that the findings could not prove the existence of predictive bias while the conclusions accept these findings as proof that there is no bias in the test. Absence of evidence is not evidence of absence as any researcher knows very well. An additional problem with the studies about bias in prediction is that some assumptions are made which may not correspond to reality, especially when logical reasonings are applied to psychometric models which have their own assumptions and their own limitations.

PURPOSE OF THE PRESENT STUDY

We propose to develop a theoretical framework to study bias in order to understand the findings and evaluate the conclusions about the research on predictive bias in the test.

Concept of Bias

Bias is a response to an item elicited by some trait on which two clearly identifiable groups differ markedly but which is not related to the trait intended to be measured by the test. Jensen (1980) used a very similar definition of bias among the many definitions he proposed.

Bias is not a property of the item but rather it is a relationship between the response and the trait; the meaning of a Russian word would be a bias item if included in a test of intelligence but it would be quite appropriate if it is included in a test of the Russian language. Bias does not indicate that the response is elicited exclusively by other trait; it is unconceivable that an item that obviously does not measure an intended trait was included in the test. A biased item has to measure some trait on which two groups differ markedly; the item has to be discrepant, aberrant, an outsider, showing patterns of responses different from the responses from other items.

Assumptions

During the theoretical discussion, the following assumptions are made: (1) two groups have clearly been identified as being different in some relevant characteristic of interest; (2) the two groups are different in the ability (trait) we want to measure and in the performance we want to predict; (3) each group has identical

standard deviation on the predictor and on the criterion, in order to simplify the analysis, indicating that the slopes of the regression lines are given by the correlation coefficients; (4) the two groups have the same validity coefficient or correlation between predictor and criterion and the same intercept in the regression lines. Under these assumptions and before bias is introduced in the test, the regression lines for both groups coincide and the difference between the groups would be in their predictor and criterion means.

If bias is introduced in the test, we make two additional assumptions. First, none of the persons in the low scoring group answers correctly any of the biased items; this is an unrealistic assumption because even if the items are completely unknown there is always some guessing in a multiple choice test. Second, we also assume that the scores of the high scoring group on the biased part of the test are unrelated both to the scores on the unbiased part and to the scores on the criterion; the consequence of this assumption is that the correlations between the unbiased test, the biased test, and the criterion are zero. Under these assumptions, bias affects only the parameters of the predictor in the high scoring group and the corresponding regression line. The low scoring group has the same parameters in the biased than in the unbiased test.

In the consideration of predictive bias we depart drastically from the treatment the topic has received in the literature in one important respect. We assume that the two groups are different in the aptitude measured by the test because the hypothesis of identical aptitude among groups is very unrealistic. The overwhelming empirical evidence has shown a mean difference in aptitude between the majority and minority groups, usually blacks, of about one standard deviation; this finding has been quite consistent across different studies, different tests, and different situations. Given the differences in sociological and educational conditions between blacks and whites that exist and have existed for so long, considering that there is no fundamental distinction between achievement and aptitude tests, given the fact that test scores differences are very large and consistent, it is very hard to

assume that the two groups are identical in aptitude. The position that " all observed differences in mental test scores are due to a built-in cultural bias of the tests themselves" (Reynolds & Brown, 1984, p.VII) is too broad to test the hypothesis of test bias and we consider as more acceptable the hypothesis that scores on psychological tests may be accurate estimates of the ability of whites but they are poor estimates of the ability of minority persons (Hunter et Al., 1984). Our hypothesis, therefore, is that test scores may exaggerate the real differences in ability between minority and majority groups.

BIAS EFFECT ON GROUP PARAMETERS

Under our assumptions, bias does not affect any parameter of the low scoring group, and in the high scoring group it affects the parameters of the predictor and of the regression line. We present the formulas of the new parameters in relation to the parameters before bias was introduced; a complete derivation of the formulas is presented in the Appendix. We also discuss how the changes in each parameter may affect the comparisons between the two groups and the interpretation of empirical findings about differential validity and differential prediction.

Predictor Mean

The high scoring group mean in the total test, biased and unbiased items, will be $\overline{X} + \overline{a}$, the mean of the unbiased part ($\overline{X}$) plus the mean of the biased items ($\overline{a}$). The bias in the test always , without exception, increases the mean predictor score of the high scoring group because $\overline{a}$ is always positive. As a result, the difference in ability between two groups is proportionally increased by the amount of bias present in the test.

This obvious and always present effect has not received, however, much attnetion in the literature and the two main reason have been either that the mean is not suitable for the detection of bias (Jensen, 1980) because it may reflect true differences between the groups or that the difference in average group performance does not reflect bias in the instrument (Wigdord & Garner, 1982). Average group differences, even large group differences in a test, do not indicate bias by themselves, but we can also say that (1) group differences in a test do not reflect by themselves actual differences on the trait and (2) if there is bias in the test, it will most likely occur where large average group differences have been found.

The controversy about test bias originated because substantial differences

were discovered between groups and for some people the differences were unreasonably large to be explained by any conceivable theory of aptitude development. This does not mean sponsoring the egalitarian hypothesis that groups are identical in any psychological characteristic; it simply implies that group differences are expected given the particular socioeconomic and educational characteristics of the two groups compared; but extremely large group differences make the instruments questionable.

Predictor Variance

The high scoring groups predictor variance in the total test is $\sigma_p^2 = \sigma_x^2 + \sigma_a^2$ under the assumption of no relationship between the biased and the unbiased parts of the test. As a general rule, bias in a test increases the variance of the test. The variance of a biased test can never be smaller than the variance of a debiased test but it can be the same if the biased items are answered correctly by every individual in the group, in which case σ_a^2 will be zero. The importance of bias in this parameter will be seen when we analyse the validity and the slope of the regression line, parameters that are directly affected by changes in the variance of the test.

Validity Coefficient

The validity coefficient for the high scoring group, correlation between predictor and criterion, after bias is introduced in the test is given by the formula $r_{(x+a)y} = r_{xy}(\sigma_x/\sigma_p)$. Being σ_p never smaller and usually greater than σ_x, the validity of a biased test can never be greater than the validity of a debiased test but it can be the same if $\sigma_p = \sigma_x$ or when there is no increase in the standard deviation as a result of the bias introduced in the test. Bias, under the assumptions we made, usually lowers the validity of a test.

The reduction in validity, however, is not directly related to the amount but rather to the variability of bias. If the biased part is completelt biased (every

single individual answers correctly every biased item) there is no change in the variability between biased and unbiased scores and the validity coefficient will not change. The largest change in variability and the largest reduction in validity takes place when there is moderate amounts of bias rather than when there is too much bias. Being this the situation, we believe that the validity coefficient is the wrong parameter to detect bias.

These considerations have important implications in the discussion of differential validity. The overwhelming research findings have been that there are not significant differences between the validity coefficient for high and low scoring groups in most tests used in selection, placement, and college admissions. We can not conclude, however, that these findings demonstrate that the test is not biased or that the issue of differential validity is baseless.

With respect to bias, there is the common impression that bias lowers the validity for minorities for not for the majority. If we define bias as that part of a test which is not related to the trait nor to the criterion but on which two groups differ markedly, bias most likely lowers more the validity coefficient of the high than of the low scoring group depending on the changes in the standard deviation for each group. A non-significant difference in validity between two groups could indicate that there is no bias in the test or that there is too much bias because in both cases bias does not introduce substantial changes in the validity coefficient.

With respect to the differential validity hypothesis, it assumes that the trait measured by the test may be better predictor of performance in the high than in the low scoring group, and this issue can only be studied if the test is not biased. If there is bias in the test, it usually lowers the validity coefficient and the rejection of the differential validity hypothesis could indicate that the test is equally valid because of bias and without bias the test would be more valid for the high than for the low scoring group. Bias and differential validity are different issues and, although related in a particular way, they have to be studied as separate issues.

Slope of the Regression Line

After the introduction of bias in the predictor, the slope of the regression line for the high scoring group is given by $b' = b_{(x+a)y} = b_{xy}(\sigma_x^2 / \sigma_p^2)$ indicating that bias reduces the slope by an amount proportional to the increase in the variance of the test. The slope of a biased test, under our assumptions, can never be greater than the slope of a debiased test.

The slope represents the relationship between changes in the criterion as a result of the changes in the predictor. Every other thing being equal, the lower the slope the smaller the predicted criterion scores. A biased test usually underpredicts the criterion scores of high scoring groups because bias lowers the slope of the regression line. This effect is contrary to the belief that bias in a test will overpredict, give an unfair advantage, the criterion performance of the group favored by the bias part. We should not, however, jump to the conclusion that bias is against instead of in favor of the high scoring group, an issue that has to be resolved through a complete, rather than a partial, analysis of the regression line. Every other thing being equal, bias underpredicts the criterion scores for the members of the high scoring group.

The slope is not the proper parameter to study bias nor is it adequate to study differential prediction unless the test is unbiased. No difference in slopes between two groups could indicate (1) equal prediction if the test is not biased (2) lack of bias if the test is equally valid, and (3) both differential validity and bias operating together but in different direction and the net result would be unnoticeable.

Intercept of the Regression Line

The intercepts of the unbiased and biased lines for the high scoring group

are $\bar{Y} - b\bar{X}$ and $\bar{Y} - b'(\bar{X} + \bar{a})$ respectively. The biased will be higher than the unbiased intercept if $b'(\bar{X} + \bar{a})$ is smaller than $b\bar{X}$ or if

$$b'/b < \bar{X}/(\bar{X} + \bar{a})$$

and substituiting b' for its value of $b(\sigma_x^2/\sigma_p^2)$ we obtain

$$\sigma_x^2/\sigma_p^2 < \bar{X}/(\bar{X} + \bar{a})$$

If the ratio of the unbiased to the biased variance is smaller than the ratio of the nbiased to the biased mean, then the biased intercept is higher than the unbiased intercept and the biased line will be positioned above the unbiased line.

Table 1 reports the values of the biased standard deviation corresponding to

---

Table 1 about here

---

to different values of the biased mean necessary for the biased line to be positioned above the unbiased line.

The numbers in this table indicate that as long as the increase in the predictor mean due to bias in the test is twice as large as the increase in the standard deviation the intercepts of the biased and unbiased lines will be the same. If the increase in the mean is more than twice the increase in the standard deviation, then the biased intercept will be larger than the unbiased intercept and the biased regression line will be above the unbiased regression line. If bias increases the mean by 20% and the standard deviation increases by more than 10% the biased is above the unbiased regression line.

The effect of bias on the intercept of the regression line is a complex relationship between changes in the slope and changes in tne predictor mean. Equal intercepts do not necessarily indicate unbiasness in the test and it could be the result of the combined effect of changes in the mean and on the slope. Because there is no direct linkage between bias and intercept, equality of intercepts can be studied only after the test has been debiased.

Crossing of the Biased and Unbiased Lines

The biased and unbiased lines cross each other at a point where the ordinates and the abscissae are the same, or $bX + C = b'X + C'$ and $X = (C - C')/(b - b')$. This point is always below the unbiased mean for the high scoring group. Substituiting the intercepts for their values we have $X = \overline{X} - (b'-\overline{a})/(b-b')$ and, because the denominator is always positive if there is bias in the test, the value of X is smaller than $\overline{X}$ and the crossing of the two lines takes place at a point below the unbiased mean for the high scoring group.

The crossing wil be above the mean for the low scoring group if $(X_H - X_L) > (ab')/(b-b')$. If the difference between the unbiased means for the two groups is large, the amount of bias and the reduction in the slope are small, then the crossing of the two lines will take place at a point above the mean for the low scoring group.

DISCUSSION OF THE THEORETICAL EFFECTS OF BIAS

The overall effects of bias are graphically represented in Figure 1. Line A is the unbiased line common to both the low and the high scoring group; the only difference between the two groups is on the predictor and on the criterion means. Line B represents the biased regression line for the high scoring group; the low scoring group does not have a biased line because bias, under our assumptions, does not change any of the parameters of the group.

Regression line B is a realistic representation of the effects of bias on the high scoring group: 1) bias increases the mean predictor scores, 2) lowers the slope, 3) it probably also raises the value of the intercept, and 4) the crossing between the biased and the unbiased liens is always below the unbiased mean for the high scoring group and very likely it is above the mean for the low scoring group if both groups differ markedly on the unbiased predictor.

Effects of Bias on the High Scoring Group

Bias produces the following results in the high scoring group:

1. The predictor mean increases by the average amount of bias in the test and, in that respect, bias favors the high scoring group.
2. The mean criterion score is underpredicted. The high scoring group, as an average, would have a higher predicted criterion mean corresponding to the biased predictor mean if we used the unbiased than if we used the biased line.
3. The biased regression line underpredicts the criterion scores of most individuals in the high scoring group. Beyond the crossing of the biased and the unbiased lines (the great majority of the subjects), for the same predictor score the predicted criterion would be lower using the biased than if we

used the unbiased line.

4. The biased regression line also overpredicts the criterion scores of some of the individuals in the high scoring group, all those subjects scoring below the crossing of the biased and the unbiased lines.
5. There are, therefore, under- and overpredictions but there are not systematic over- or underpredicted criterion scores.

The impression from these conclusions that the biased regression line is disadvantageous to the high scoring group as a whole because it underpredicts their mean criterion in relation to the prediction made by the unbiased line would be erroneous. It would also be erroneous to assume that the biased line is advantageous to the low scoring group because, as we will discuss later, it overpredicts the group's mean criterion score. Analysing Figure 1 we can see that the predicted criterion mean corresponding to the high scoring group's biased mean is larger when it is predicted from the unbiased than when the prediction is made from the biased line and the opposite is the case with the low scoring group where the predicted mean is smaller in the unbiased than in the biased line.

The only time when there are actual criterion and predictor scores is at the time of validating the test and this should be the moment of studying the issue of advantage or disadvantage to any one of the groups. There are three means for the group on which the test is validated: the actual criterion mean, the actual predictor mean on the biased test, and the actual predictor mean on the debiased test. Bias increases the mean predictor score of the high scoring group but it does not affect in any way the actual criterion mean. As a result, the same criterion mean corresponds to a lower predictor mean in the unbiased than in the biased line. We can analyse this situation in two ways; we can say that the same predictor mean has a higher predicted criterion mean using the unbiased than using the biased line and, in that sense, the biased line underpredicts the criterion mean; or we could also say that the same criterion mean corresponds to a higher predictor mean using the biased than using the unbiased line and, in that sense, the biased predictor mean is in-

flated because some part of it does not corresponds to any part of the criterion mean. Underprediction of the criterion mean by the biased line is, analytically, the same as an inflated predictor mean but the meaning conveyed by each analysis is not the same because underprediction implies disadvantage while inflated mean implies advantage. What bias actually does is to inflate the predictor mean rather than underpredict the criterion mean and bias does not represent any disadvantage to the high scoring group; on the contrary, it is an advantage when selection decissions are made, as they actually are, based on the predictor scores which are inflated by bias.

Bias does not systematically over- or underpredicts the criterion scores of the high scoring group. As long as the biased and the unbiased lines cross each other, and considering the assumptions in our study of bias they most likely do, it is very unlikely that the biased regression line systematically over- or underpredicts the criterion scores of the majority or of any other group and systematic over- or underprediction is a very unlikely empirical event. In our analysis, the two lines cross each other because bias usually lowers the slope and increases the intercept of the regression line.

Effects of Bias on the Low Scoring Group

Using the biased regression line for predictions in the low scoring group we come out with the following results:

1. The biased regression line overpredicts the mean criterion score and most of the individual scores for the low scoring group. All those subjects with a predictor score below the point where the biased and the unbiased lines cross each other (more than 50% of the subjects) will receive an overpredicted criterion score.
2. There are also underpredicted criterion scores for all those individuals scoring on the predictor below the crossing of the two lines (less than 50% of the subjects in the group).

3. The biased regression line does not systematically over- or underpredicts the criterion scores of the low scoring group because we obtained over- and underpredictions rather than predictions in only one direction.

Main Effect of Bias

These are not the only results of bias in a test and the main effect is a shift in the position of the two populations because of the changes in the predictor means brought about by the amount of bias in the test. The overlapping of the two populations is substatially reduced by the introduction of bias and, if the test is used for selection and only high scorers are chosen, the number of selectees from the low in relation to the high scoring group will be reduced by bias and the proportion would be smaller the higher the predictor qualifying score is. Suppose that the real difference in predictor means between two groups is one standard deviation. Depending on the amount of bias, as reported in table 2, the number of selected candidates from each group will change drastically.

Table 2 about here

If the qualifying score is the mean predictor score for the high scoring group, bias does not produce any unfavorable effect on the number of selectees from the high scoring group because, regardless of the amount of bias, 50% of the individuals would always be selected. Important reduction in the number of selectees take place in the low scoring group, from 16% who reached the qualifying score when there is no bias to only 2% when bias amounts to one standard deviation.

If the qualifying score is kept at the level of the mean for the high scoring group on the unbiased test, no changes take place because of bias in the low scoring group but important increases of qualifying individuals occur in the high scoring group. About 16% qualify from the low scoring group in both tests, but the percentage for the high scoring group goes from 50% when there is no bias, to 69% when the amount

of bias is .50 SD, to 93% when the amount of bias is larger.

Regardless of the way we look at the situation, the low scoring group always gets a rough treatment from bias and the high scoring group is always favored by bias. The overprediction of criterion scores for the low scoring group represents no advantage because fewer and fewer of the individuals are going to be selected as the amount of bias increases. The underprediction of the high scoring group is not a disadvantage when more and more individuals are going to be selected as the amount of bias increases.

The larger the amount of bias present in the test, the larger the overprediction for the low scoring group but the smaller the number of individuals to be selected, the larger the underprediction for the high scoring group but the larger the number of individuals to be selected. If we want to make the largest overprediction as a way of given an advantage to the low scoring group, the lesson would be to make the test completely biased; but then, there is little to be preddicted because none of the individuals would be selected. If we want to make the largest underprediction as a ways of taking away the advantage of the high scoring group, the lesson would be to make the test more and more biased; but then, fewer and fewer would be deprived of anything because larger and larger number of individuals would be selected as the amount of bias increases.

OVERPREDICTION AND TEST BIAS

Empirical data have shown that the common regression line overpredicts the mean and most of the individual scores for minority groups and underpredicts the mean and most of the individual scores for the majority. Assuming that overprediction means favoring, the general published conclusions from the research studies with the regression line have been that tests do not adversely discriminate against minorities and the tests are, therefore, unbiased (Cleary, 1968; Petersen & Novick, 1976; Jensen, 1980, 1984; Anastasi, 1982; Hunter et Al., 1984). Hunter et Al. (1984) present a clear rational argument involved in drawing these conclusions. If the majority and minority groups are equal in ability and if there is bias in the test, the minor group regression line has to be positioned above the major group regression line in which case minority subjects would have consistently higher predicted criterion scores than subjects from the majority group with the same predictor scores. Figure 2A would be the graphic representation of this theoretical case of bias in prediction; empirical data, as represented in Figure 2B, have shown the opposite, that is, the minor group regression line is below the regression line for the major group.

One proble with this and other similar reasonings is that the hypothesis to be tested, the position of the two regression lines, is based on two assumptions: the two groups are equal in ability and the test is biased. The rejection of the hypothesis could be the result of being wrong both or either one of the assumptions. If it was wrongly assumed that the two groups were equal, as the overwhelming evidence from empirical data has shown, then the empirical data neither support nor contradict any conclusion about test bias. The discussion throughout this paper has been based on the assumption of unequal ability among the groups and the most likely effect of bias in the test is the positioning of the major group regression line above the regression line for the minor group, the same position shown by empirical data.

There is also the danger of making the wrong conclusion when a rational argument is used to predict the behavior of a very well defined and limited psychometric model.

The regression line does not represent a common sense proposition but rather it is a very limited psychometric model under certain conditions and in order to accomplish a very specific result. Assuming that the conditions of normality and homoscedasticity for a linear relationship are met, the purpose of the linear regression line is merely to predict criterion scores in such a way that the errors of prediction will be reduced to a minimum and in that case the regression line becomes the best fit line under the least square condition. The regression line, like any other statistic in the classical test theory, is group dependent and it, therefore, has to be used very carefully when comparing groups which are different in ability. The regression equations for two different groups ($\overline{X}_H \neq \overline{X}_L$), each one of them having the same mean and standard deviation on the predictor and on the criterion are given by

$$Y_H = r.X + (1 - r)\overline{X}_H$$

$$Y_L = r.X + (1 - r)\overline{X}_L$$

where r is the validity of the test and H and L stand for high and low scoring group.

If the validity is perfect (r = 1), then the predicted criterion scores are identical to the predictor scores and they are the same in the high as in the low scoring group; an individual will obtain the same predicted criterion score regardless of group membership. But if the validity is less than one, group membership becomes a factor in determining the predicted criterion score an individual will get and group membership becomes more relevant the lower the validity coefficient and the larger the groups mean difference are. If the standard deviation is 16 in both groups and their means differ by one standard deviation ($\overline{X}_H$ = 100 and $\overline{X}_L$ = 84), an individual with a predictor score of 90 will obtain the predicted criterion scores shown in table 3 depending on the validity of the test and on the group he belongs to.

Table 3 about here

The numbers in this table make many people unconfortable because it is against our common sense to accept, even though it is psychometrically correct, that the same person can have two different predicted scores depending on the group to which, by

accident or by nature, he belongs to and the difference between the two predicted scores could be substantial (about .70 SD when the validity is .30). Goodenough & Harris (1952) reminded us that tests are good at ranking individuals within a very homogeneous group but they are not good at ranking heterogeneous groups. The group dependency of predicted scores presents a problem for the discussion of predicted test bias if the groups are different in ability.

Group Advantage

Over- and underprediction have been interpreted as favoring or being against a particular group or individual but this is only one definition. Favoring could also mean the result of a comparison of either the predictor or the predicted criterion scores with some other standard which in selection could be the actual criterion scores of individuals or groups. This comparison can be done at the time of conducting the validation study when all the subjects have a predictor score, a criterion score, and a predicted criterion score that can be calculated from the regression line. The best standard for comparison is actual performance because the purpose of selection and the use of tests as selection devices is to insure that the best performers are chosen and the best way of determining the usefulness of a test in future selection decisions is to analyse the available data for the group involved in the validation study.

The Predicted Scores Favor High Scorers

We can base our selection decision on the predicted performance calculated from the actual predictor scores. Advantage in that case would be a predicted criterion score that is higher than the actual performance. Using the predicted instead of the actual performance we make an error of prediction which is positively related to the predicted scores and the higher the predicted scores the larger the error of prediction. If we define as error of prediction the difference between actual and predicted

criterion scores ($e = Y - Y'$), then the correlation between error of prediction and predicted scores is always positive and, as shown in the Appendix, it is given by

$$r_{ey} = \sqrt{1 - r_{xy}^2}$$

and the regression of error of prediction over actual criterion scores would be

$$e = (1 - r_{xy}^2)(Y - \bar{Y})$$

The error of prediction is positively related to the actual performance scores and the correlation coefficient is higher the lower the validity of the test is. The higher the actual performance the larger the error of prediction; the smaller the validity of the test the larger the correlation between error of prediction and actual per ormance and the larger also the errror of prediction for any individual score. From this initial analysis it is apparent that the errors of prediction favor more the high than the low scoring individuals and groups.

In the regression equation we can observe that the error of prediction for an individual score (1) is indirectly related to the validity of the test; the lower the validity the larger the value of $(1-r^2)$ and the larger will also be the error of any particular score and (2) is directly related to the departure of the actual criterion scores from the group men $(Y - \bar{Y})$, and the error of prediction is positive for performance scores above the mean and negative for performance scores below the mean. The net effect of these positive and negative errors of prediction on the total group is a minimum because they cancell each other; the regression line does what we required it to do, that is, to minimize the errors of prediction.

The invalidity of the test, the same as the reliability, introduces errors in the prediction and these errors favor the high scoring individuals and groups in two ways. The high scoring individuals and groups predicted performance scores have positive errors (are inflated) and the low scoring individuals and groups predicted performance scores have negative errors of prediction (are deflated). The actual difference in performance between the two groups is doubly augmented by inflating high scorers predicted scores and deflating low scorers predicted scores. As a result,

the invalidity of the test and the prediction clearly favor the high scoring individuals and groups.

Predictor Scores Favor High Scorers

A similar analysis can be done comparing the scores on the predictor and the actual performance scores and the result of this analysis will be that predictor scores favor high scorers individuals and groups and the advantage is larger the lower the validity of the test is.

Using the predictor instead of the actual criterion scores to make the selection we introduce some error, error of selection to differentiate it from the error of prediction, and it would be $e = X - Y$, and the regression of this error, as shown in the Appendix, is given by

$$e = (1 - r_{xy})(X - \overline{X})$$

This formula is similar to all the formulas that try to relate errors due to unreliability or invalidity with actual or predicted scores. The error introduced by using the predictor instead of the actual performance scores is positively related to the predictor scores and the errors are positive for scores above the group's mean and negative for scores below the mean. The difference between high and low scoring individuals or groups is doubly enlarged by any error that we make in the selection which, under the condition of minimizing the errors through the least square fit mathematical formula, always inflates high scorers scores and deflates low scorers scores with the total aggregate effect being a minimum. Contrary to what Jensen (1980) said, an analysis of the errors made in selection leads us to the conclusion that invalidity in a test favors high scoring individuals and groups.

Unbiasness is not the Same as Equal Prediction

Suppose that we have two groups with identical mean difference on the predictor and on the criterion and the low scoring group's means both on the criterion and on

the predictor are one standard deviation below the means for the high scoring group. We can assume that in this particular situation the test is unbiased in the sense that group differences on the predictor reflect the same amount of group differences on the criterion. The two regression lines, however, would not be identical but rather parallel to each other and the predicted differences will be inversely related to the validity of the test as shown in table 4 where the values of the intercepts are

---

Table 4 about here

---

reported for two groups having a mean score of $\overline{X}_H = 100$ and $\overline{X}_L = 84$ both on the predictor and on the criterion. The intercept is given by

$$C = \overline{Y} - r.\overline{X} = (1 - r)\overline{X}$$

and the representation of these lines is shown in fiure 3.

Figure 3 about here

Even though the test was logically unbiased, we believe that the psychometric prediction is in favor of the high scoring group. Every single member of the high scoring group has a higher predicted score than every single member of the low scoring group with the same predictor score. The major group regression line, however, overpredicts the criterion scores of every member of the minority group; if we take overprediction as indicating favoring a particular group, the major group regression line favors the minor group. The overprediction of the low scoring group is simply an artifact of the definition of the best regression line which depends on the means of the groups and overprediction occurs even when the test is unbiased because mean group differences on the predictor and on the criterion are of the same magnitude.

Biasness is not the Same as Overprediction

A test could be clearly biased and the regression line could not detect this bias in the sense that the major group regression line could be above the regression

for the minor group. Imagine a case where the group difference on the predictor is two standard deviations but it is only one standard deviation on the criterion as shown in table 5.

Table 5 about here

In this particular case the minor group regression line should logically be above the major group regression line because group differences are smaller on the criterion (one SD) than on the predictor (two SD). Psychometrically, however, the major group is above the minor group regression line if the validity is low. The intercepts for the groups of table 5 and for different validities are reported in table 6

Table 6 about here

---

and the graphic representation is shown in figure 4.

Figure 4 about here

Two groups may differ two standard deviations on the predictor and only one standard deviation on the criterion, a clear indication of test bias against the minor group, and the minor group regression line can be, nevertheless, below the major group regression line indicating that, psychometrically, the test is not biased because minor group criterion scores would be overpredicted by the major group regression line.

In order for the minor group regression line to be above the major group regression line the minor group intercept must be larger than the major group intercept. This means that

$$\bar{Y}_L - r.\bar{X}_L > \bar{Y}_H - r.\bar{X}_h$$

and transposing terms we obtain

$$\bar{Y}_H - \bar{Y}_L < r(\bar{X}_H - \bar{X}_L)$$

If the validity of the test is .40, this formula indicates that, in order for

the minor group regression line to be above the major group regression line and being able to detect bias in the test, the difference in the criterion means has to be only 40% of what the mean difference is on the predictor. If the mean difference on the predictor is one standard deviation, in order for the minor group regression line to be above the major group regression line, the difference on the criterion means has to be (1) less than .50 SD if the validity is .50, (2) less than .30 SD if the validity is .30, and (3) less than .25 SD if the validity is .25. It is very hard to define these situations as logically fair but, unless the differences in mean criterion scores are substatially lower than the differences in predictor means, the major group regression line will overpredict the minor group criterion scores and, psychometrically, the test would be fair for minorities.

The regression line, when used with groups which differ in ability, can only detect extreme bias in the test but it can not detect moderate or even substantial bias if the validity is low. It is logically difficult to accept, even thoug it is psychometrically correct, that a test that shows group mean differences of two standard deviations is not biased unless the mean group difference on the criterion is less than .80 SD if the validity of the test is the commonly found coefficient of .40. If only 2% of the minority group score on the predictor as high as 50% of the majority, but 33% score on the criterion as high as 50% of the majority the regression line will overpredict the minority criterion scores if the validity is only .25 and the test would be psychometrically unbiased because the minor group regression line will be below the major group regression line. If the selection is based on predictor scores, as it is the case in real life, we do not see how the test will favor minorities when the validity is less than perfect and especially when it is low.

CONCLUSION

The consistent empirical findings of equal or comparable validity of tests among clearly identifieble different groups neither support nor contradict the existence of bias in the test or the hypothesis of differential validity. The validity coefficient is not a proper indicator of bias unless the hypothesis of differential validity has been rejected in advance nor is it suitable to study differential validity unless the test is unbiased.

The effects of bias in a test are an unquestionable increase in the average group differences and a probable reduction in all the other parameters with the exception of the regression intercept which is very likely increased by bias in the test. The elimination of bias should, therefore, produce a reduction in mean group differences with no decreases in the other parameters. The proposed item bias techniques have generally reported no significant changes in mean group differences when the detected biased items were eliminated from the test, suggesting that the definitions of bias, on which the techniques were based, might not be the most adequate.

The concept of overprediction of minority criterion scores as an indication of predictive bias is psychometrically acceptable only if the groups are equal in ability. If the groups possess different degree of ability, overprediction could be an indicator of extreme bias in the test but it would not detect moderate or even substantial bias if the validity of the test is low. Equal prediction, does not indicate, psychometrically, absence of bias nor does differential prediction indicate presence of bias unless we define bias this way.

The theoretical development of bias in prediction could be a useful framework to clarify some of the issues involved in the controversy about ability testing. The conclusions made from empirical data, which we reported at the beginning of the discussion, are not supported by a theoretical model of the effects of bias on the predictive equation parameters.

As a result of these considerations, we believe that the main issues of the controversy about ability testing are still waiting to be settled through different kinds of research from the ones undertaken in predictive validity studies. The main issue, as we see it, is bias in the test and no other issue like differential validity of differential prediction can be addressed until bias has been convincingly detected and eliminated from the test.

APENDIX

Biased Test Parameters

We present the derivation of the statistical formulas used during the discussion. These formulas are for the biased test.

1. Mean $\sum(X_2+a)/n = \bar{X}_2+\bar{a}$ where X_2 represents the scores on the unbiased part and a the individual scores (also a variable) on the biased part.

2. Standard Deviation

$$\sigma_p^2 = \frac{\sum(X+a)^2 - \left[\sum(X+a)\right]^2/n}{n} =$$

$$= \frac{\sum X^2 + \sum a^2 + 2\sum Xa - \left[(\sum X)^2 + (\sum a)^2 + 2\sum X \sum a\right]/n}{n} =$$

$$= \frac{\sum X^2 - (\sum X)^2/n}{n} + \frac{\sum a^2 - (\sum a)^2/n}{n} + 2\frac{\sum Xa - (\sum X \sum a)/n}{n} =$$

$$= \sigma_x^2 + \sigma_a^2 + 2r_{xa}\sigma_x\sigma_a \qquad \text{but we assumed that } r_{xa} = 0$$

$$\sigma_p = \sqrt{\sigma_x^2 + \sigma_a^2}$$

3. Validity Coefficient

$$r_{(x+a)y} = \frac{\sum(X+a)Y - \left[\sum(X+a)\sum Y\right]/n}{n\sigma_{(x+a)}\sigma_y} =$$

$$= \frac{\sum XY + \sum aY - \left[\sum X \sum Y + \sum a \sum Y\right]/n}{n\sigma_p\sigma_y} =$$

$$= \frac{\sum XY - \left[\sum X \sum Y\right]/n}{n\sigma_p\sigma_y} + \frac{\sum aY - \left[\sum a \sum Y\right]/n}{n\sigma_p\sigma_y}$$

but $r_{ay} = 0 = \dfrac{\sum aY - (\sum a \sum Y)/n}{n\sigma_a\sigma_y}$ and $\sum aY - (\sum a \sum Y)/n = 0$

$$r_{(x+a)y} = \frac{\Sigma XY - (\Sigma X \Sigma Y)/n}{n\sigma_p\sigma_y} \cdot (\sigma_x/\sigma_x) = r_{xy}(\sigma_x/\sigma_p)$$

4. Slope of the Regression Line

$$b' = b_{(x+a)y} = r_{(x+a)y}(\sigma_y/\sigma_p) = r_{xy}(\sigma_x/\sigma_p).(\sigma_y/\sigma_p) =$$

$$= r_{xy}(\sigma_y/\sigma_x).(\sigma_x/\sigma_p)^2 = b_{xy}(\sigma_x/\sigma_p)^2$$

5. Crossing of the Biased and the Unbiased Lines

The two lines cross each other at a point where the ordinates and the abscissae are the same in both lines

$$\left.\begin{aligned} Y &= bX + C \\ Y' &= b'X + C' \end{aligned}\right\} \qquad bX + C = b'X + C'$$

$$X(b - b') = (C' - C) \quad \text{and} \quad X = (C' - C)/(b - b')$$

Substituiting the intercepts for their values

$$X = \frac{\overline{Y} - b'(\overline{X} + \overline{a}) - \overline{Y} + b\overline{X}}{b - b'} = \frac{\overline{X}(b - b') - b'\overline{a}}{b - b'} = \overline{X} - \left[b'/(b-b')\right]\overline{a}$$

Error of Prediction Parameters

1. Standard Deviation

If we define the error of prediction as the difference between actual and predicted criterion scores ($e = Y - Y'$), and if we are working with deviation instead of raw scores, the variance of the errrors of prediction would be

$$S_e^2 = (1/n)\sum(y - y')^2 = (1/n)\sum(y - b.x)^2 = S_y^2 + b^2S_x^2 - 2.b(\Sigma xy/n)$$

and knowing that

$$\Sigma xy/n = S_x S_y r_{xy} \qquad \text{and} \qquad b_{xy} = r_{xy}(S_y/S_x)$$

and making these substitutions we obtain

$$S_e^2 = S_y^2 + r_{xy}^2 S_y^2 - 2.r_{xy}^2 S_y = S_y^2 - r^2 S_y^2 = S_y^2 (1 - r^2)$$

and

$$S_e = S_y.\sqrt{1 - r^2}$$

2.Correlation between Error of Prediction and Criterion

Using the definition of correlation for deviations scores

$$r_{ey} = \Sigma ey/(n.S_e.S_y) = 1/(nS_e S_y)\left[\Sigma (y - bx)y\right] = 1/(S_e S_y)\left[(S_y^2 - brS_x S_y)\right] =$$

$$= 1/(S_e S_y)\left[(S_y^2 - r^2 S_y^2)\right] = \frac{S_y^2 (1 - r^2)}{(S_y\sqrt{1 - r^2})S_y} = (1-r^2)/(\sqrt{1-r^2}) = \sqrt{1 - r^2}$$

3. Regression of Error of Prediction over Criterion Scores

Assuming that there is a linear relationship between error and criterion scores (we already know that these two variables are positively related) and using raw instead of deviation score, the regression equation would be

$$e = b_{ey}.Y + 0 - b_{ey}.\bar{Y} = b_{ey}(Y - \bar{Y})$$

and knowing that

$$b_{ey} = r_{ey}(S_e/S_y) = (\sqrt{1 - r^2})(S_y\sqrt{1 - r^2})/S_y = 1 - r^2$$

then

$$e = (1 - r^2)(Y - \bar{Y}) = (1 - r^2).y$$

Error from Using Predictor instead of Criterion

1. Standard Deviation of Error

The error we make by using the predictor instead of the criterion scores would be $e = X - Y$, the variance would be

$$S_e^2 = (1/n)\Sigma(x - y)^2 = S_x^2 + S_y^2 - 2.r.S_xS_y$$

and the square root would be the standard deviation

2. Correlation between Error and Predictor Scores

$$r_{ex} = (1/nS_eS_x)\Sigma(x - y)x = (S_x - r.S_y)/\sqrt{S_x^2 + S_y^2 - 2.r.S_xS_y}$$

and the slope of the regression line would be

$$b_{ex} = r_{ex}(S_e/S_x) = \frac{S_x - r.S_y}{\sqrt{S_x^2 + S_x^2 - 2rS_xS_y}} \cdot \frac{\sqrt{S_x^2+S_y^2 - 2rS_xS_y}}{S_x} = 1 - r(S_y/S_x)$$

and

$$b_{ex} = 1 - b_{xy}$$

and the regression line would be

$$e = (1 - b)(X - \bar{X})$$

and when $S_x = S_y$ or the scores both on the criterion and on the predictor are in standardized form, the regression equatioj becomes

$$e = (1 - r_{xy})(X - \bar{X})$$

Table 1: Proportions of the Biased to the Unbiased Standard Deviations for Different Proportions of the Biased to the Unbiased Means which will Make Identical the Biased and the Unbiased Intercepts

Biased Mean in Relation to Unbiased Mean	Biased SD in Relation to Unbiased SD
$1.10\bar{X}$	$1.05\sigma_x$
$1.15\bar{X}$	$1.07\sigma_x$
$1.20\bar{X}$	$1.10\sigma_x$
$1.30\bar{X}$	$1.14\sigma_x$
$1.40\bar{X}$	$1.18\sigma_x$
$1.50\bar{X}$	$1.22\sigma_x$

Table 2: Percentage of Selectees from High and Low Scoring Groups for Different Amounts of Bias in the Test

Amount of Bias	High Scoring Group Mean as the Selection Criterion		Unbiased High Scoring Group Mean as Selection Criterion	
	Low Scorin Group	High Scoring Group	Low Scoring Group	High Scoring Group
0.0	16%	50%	16%	50%
0.5 SD	7%	50%	16%	69%
1.0 SD	2%	50%	16%	84%
1.5 SD	1%	50%	16%	93%

N.B.: The mean group difference is one SD.

Yable 3: Predicted Scores for an Individual Having a Predictor Score of 90 depending on the Validity and Group Membership

Validity	Group Mean of 100	Group mean of 84
.60	94	87.6
.40	96	86.4
.30	97	85.8

Table 9: The Intercepts for Two Groups with Means of $X_H = Y_H = 100$ and $X_L = Y_L = 84$ and Different Values of the Validity.

r	Major Group	Minor Group	Difference
.80	20	16.8	3.2
.60	40	33.6	6.4
.40	60	50.4	9.6

Table 5: Theoretical Means and Standard Deviations for Majority and Minority Groups

	Major Group		Minor Group	
	Mean	SD	Mean	SD
Predictor	20	4	12	4
Criterion	20	4	16	4

Table 6: The Intercepts of the Regression Lines for Groups of Table 5 and for Different Validity Coefficients

Validity	Major Group	Minor Group
.40	12	11.2
.30	14	12.4
.25	15	13

Figure 1: The Biased and Unbiased Regression Lines for Low and High Scoring Groups

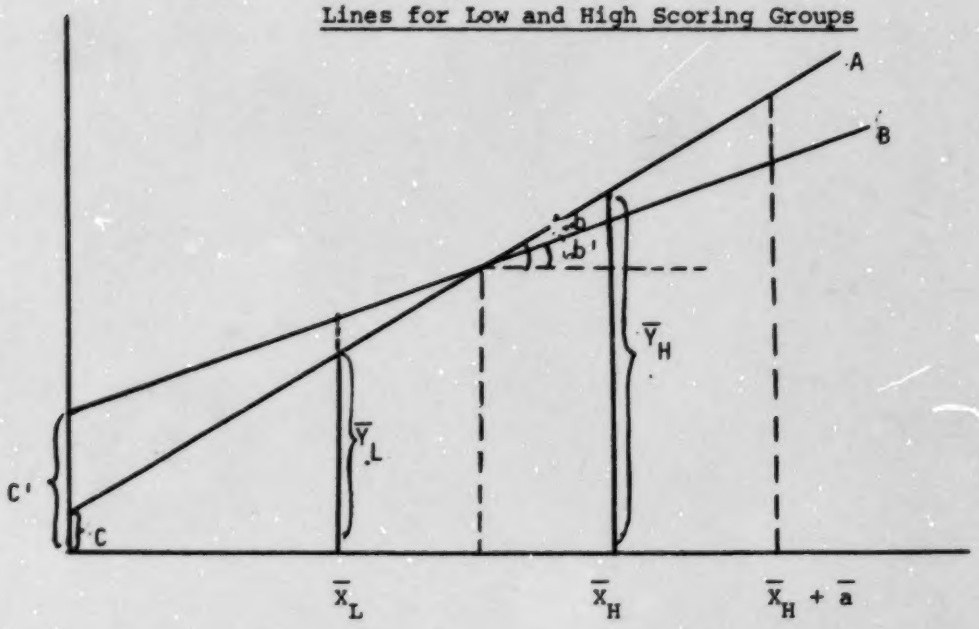


Figure 2: Graphic Representation of Theoretically Predicted Biased (A) and Actual Regression Lines(B) from Empirical Data

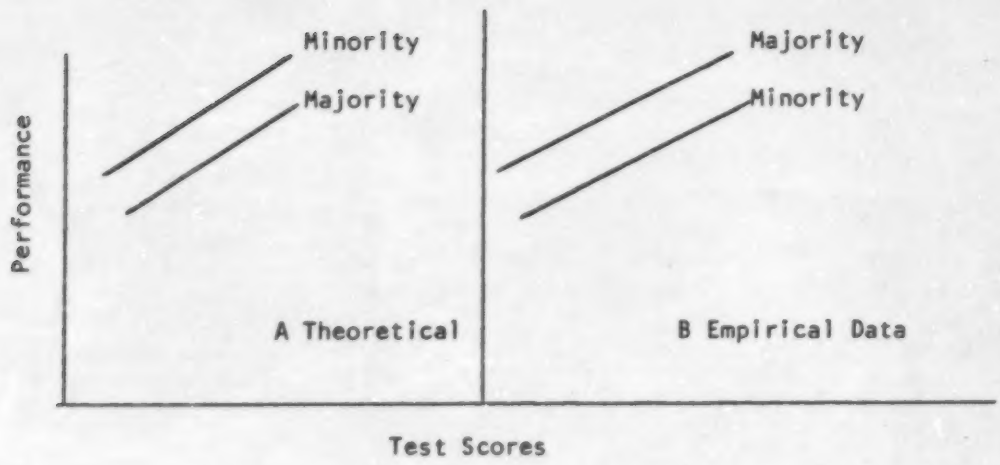


Figure 3: Graphic Representation of the Regression Lines for Two Groups with Identical Mean Differences of one SD on the Predictor and on the Criterion for Different Validity Values.

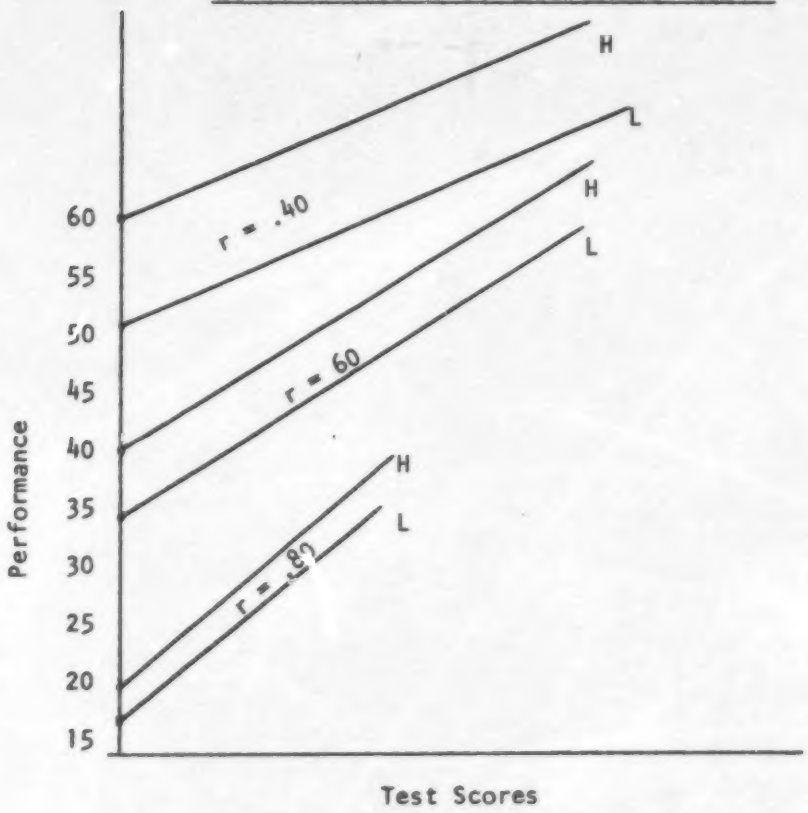
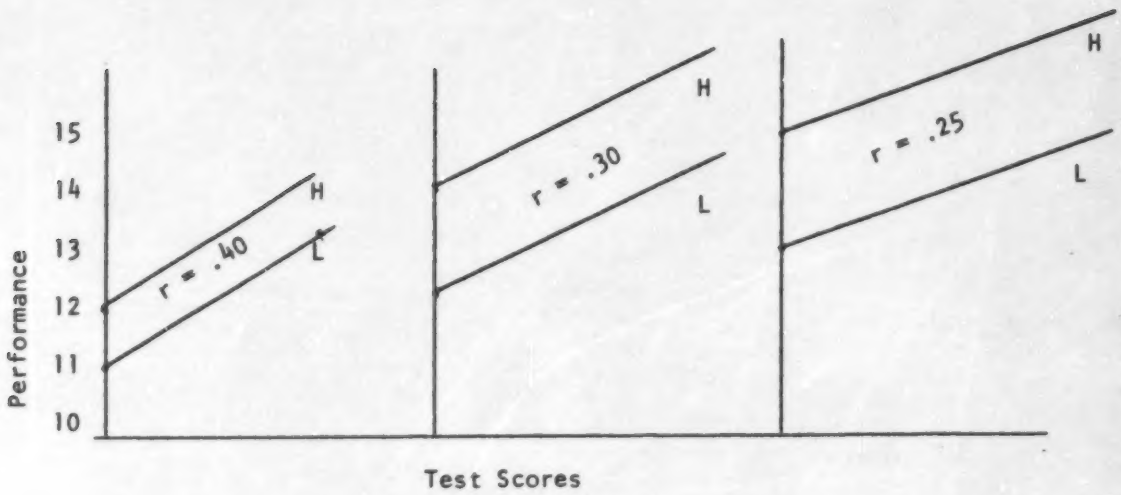


Figure 4: Graphic Representation of the Regression Lines for the Groups Described in Tables 5 and 6



BIBLIOGRAPHY

Anastasi, A. Psychological Testing (5th Ed.). McMillan Publishing Company, Inc. New York, 1982.

Boehm, V.R. Differential Prediction: A Methodological Artifact? Journal of Applied Psychology, 1979, 62, 146-154.

Cleary, T.A. Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Goodenough, F.L. & Harris, D.B. Studies in the Psychology of Children's Drawings: 1928-1949. Psychological Bulletin, 1950, 47, 369-433.

Hunter, J.E., Schmidt, F.L., & Rauschenberger, J. Methodological, Statistical, and Ethical Issues in the Study of Bias in Psychological Tests. In Reynolds, C.R. & Brown, R.T. (Eds.), Perspectives on Bias in Mental Testing. Plenum Press: New York, 1984.

Ironson, G.H. Use of Chi-square and Latent Trait Approaches for Detecting Item Bias. In Beck, R.A. (Ed.) Handbook of Methods for Detecting Test Bias. The Johns Hopkins University Press: Baltimore, Maryland, 1982.

Jensen, A.R. Bias in Mental Testing. The Free Press: New York, 1980.

Jensen, A.R. Test Bias: Concepts and Criticisms, In Reynolds, C.R. & Brown, R.T. (Eds.) Perspectives on Bias in Mental Testing. Plenum Press: New York, 1984.

Linn, R.L. Single Group Validity, Differential Validity, and Differential Prediction. Journal of Applied Psychology, 1978, 63, 507-512.

Petersen, N.S. & Novick, M.R. An Evaluation of Some Models for Cultural-fair Selection. Journal of Educational Measurement, 1976, 13, 3-29.

Reynolds, C.R. & Brown, R.T. (Eds.). Perspectives on Bias in Mental Testing. Plenum Press: New York, 1984.

Wigdor, A.H. & Garner, W.R. (Eds.). Ability Testing: Uses, Consequences, and Controversies. National Academy Press: Washington D.C., 1982.

END

DEPT. OF EDUCATION

NAT'L INSTITUTE OF EDUCATION

ERIC

DATE FILMED

J U L Y

2 - 1985